Social Contracts, Fair Play, and the Justification of Punishment

Richard Dagger*

In recent years, the counterintuitive claim that criminals consent to their own punishment has been revived by philosophers who attempt to ground the justification of punishment in some version of the social contract. In this paper, I examine three such attempts—"contractarian" essays by Christopher Morris and Claire Finkelstein and an essay by Corey Brettschneider from the rival "contractualist" camp—and I find all three unconvincing. Each attempt is plausible, I argue, but its plausibility derives not from the appeal to a social contract but from considerations of fair play. Rather than look to the social contract for a justification of punishment, I conclude, we would do better to rely on the principle of fair play.

I. INTRODUCTION

There is nothing altogether new in the claim that criminals somehow have authorized or consented to their punishment. The social contract theorists of the seventeenth and eighteenth centuries advanced such a claim, each in his own way, with Jean-Jacques Rousseau stating the point most vividly: "it is in order not to be the victim of a murderer that a person consents to die if he becomes one."[1] Rousseau's reasoning, however, strikes many readers as counterintuitive. To say that a murderer deserves to die for his or her crime is controversial in itself; to say that the murderer has *consented* to be executed seems not only to defy experience but to stretch the concept of consent further than it can reasonably go. That, presumably, is why the vigorous debates over the justification of criminal punishment in the twentieth century had so little to say about the putative connection between the offenders' consent and their punishment.[2]

---

* E. Claiborne Robins Distinguished Chair in the Liberal Arts, Department of Political Science and Program in Philosophy, Politics, Economics, and Law, University of Richmond, Richmond, VA 23173, rdagger@richmond.edu. The author thanks Christopher Bennett, Zachary Hoskins, Jeffrie Murphy, Mary Sigler, Christopher Heath Wellman, and the *OSJCL* staff for helpful comments on an earlier draft of this essay.

[1] JEAN-JACQUES ROUSSEAU, ON THE SOCIAL CONTRACT 64 (Roger D. Masters ed., Judith R. Masters trans., St. Martin's Press 1978) (1762).

[2] *See, e.g.*, Michael Davis, *Punishment Theory's Golden Half Century: A Survey of Developments from (About) 1957 to 2007*, 13 J. ETHICS 73 (2009). Davis's twenty-seven-page survey does not mention consent and refers to social contract theory only twice: once to refer to what "*would be true* of a theorist who argued that the institution of punishment is required under an actual

In recent years, though, this old and perhaps counterintuitive claim has been given new life by philosophers who believe that the justification of punishment is grounded in some version of the social contract. Whether their efforts to revive this old claim are successful is the question I address in this essay. I doubt that they are, as I shall explain below, but I also believe that these efforts at revival are worthy of close scrutiny. Among other things, the defects of these efforts will tell us something about the merits of a rival theory of punishment—that is, a theory that justifies punishment not in terms of consent or contract but of fair play. Indeed, I shall argue that whatever plausibility the contract-based theories possess is largely owing to their implicit reliance on considerations of fair play.

The argument proceeds in four stages. In the first, I relate these new efforts to link punishment to the consent of the punished to broader developments within social contract theory. Next, I examine two essays—one by Christopher Morris and one by Claire Finkelstein—that grow out of what has come to be called the *contractarian* version of social contract theory.[3] Then, in section IV, I examine Corey Brettschneider's recent attempt to justify punishment by means of the *contractualist* version of social contract theory.[4] In part V, finally, I try to show how fair-play theory captures the intuitive appeal of Morris's, Finkelstein's, and Brettschneider's approaches in a more straightforward and plausible manner than their reliance on consent and contract can do.

II. THE RATIONAL, THE REASONABLE, AND THE SOCIAL CONTRACT

Professors Morris, Finkelstein, and Brettschneider will no doubt maintain that the differences in their respective theories are at least as important as their common features, but I suspect that all three will acknowledge that their essays grow out of the broader revival of interest in social contract theory sparked by the publication of John Rawls's *A Theory of Justice*.[5] There is neither space nor need, fortunately, to rehearse the many important features of Rawls's theory here, but his conception of the social contract does require attention. On Rawls's account, as set out in *A Theory of Justice*,

> The merit of the contract terminology is that it conveys the idea that principles of justice may be conceived as principles that would be chosen by rational persons, and that in this way conceptions of justice may be

---

constitution or historical social contract," *id.* at 83 (emphasis added), and once in connection with what he calls "fairness" (and I shall call "fair-play") theory. *Id.* at 94.

[3] Christopher W. Morris, *Punishment and Loss of Moral Standing*, 21 CANADIAN J. PHIL. 53 (1991); Claire Finkelstein, *A Contractarian Approach to Punishment*, *in* THE BLACKWELL GUIDE TO THE PHILOSOPHY OF LAW AND LEGAL THEORY 207 (Martin P. Golding & William A. Edmundson eds., 2005).

[4] Corey Brettschneider, *The Rights of the Guilty: Punishment and Political Legitimacy*, 35 POL. THEORY 175 (2007).

[5] JOHN RAWLS, A THEORY OF JUSTICE (rev. ed. 1999).

> explained and justified. The theory of justice is a part, perhaps the most significant part, of the theory of rational choice.[6]

Hence Rawls's reliance on the "original position" and the "veil of ignorance" to ensure that the parties to the hypothetical contract would choose principles that would be both *rational*—that is, serving to advance the individual interests of the contractors—and *impartial*.[7]

In subsequent writings, however, Rawls developed and defended his theory, which he called "justice as fairness,"[8] as a form of *reasonable* rather than *rational* choice.[9] In his *Political Liberalism*, in a section entitled "The Reasonable and the Rational," Rawls traced this distinction to the philosophy of Immanuel Kant and, more directly, to an essay by W.M. Sibley, *The Rational Versus the Reasonable*.[10] According to Sibley,

> [I]f I desire that my conduct shall be deemed *reasonable* by someone taking the standpoint of moral judgment, I must exhibit something more than mere rationality or intelligence. To be reasonable here is to see the matter—as we commonly put it—from the other person's point of view, to discover how each will be affected by the possible alternative actions; and, moreover, not merely to "see" this (for any merely prudent person would do as much) but also to be prepared to be disinterestedly *influenced*, in reaching a decision, by the estimate of these possible results. *I must justify my conduct in terms of some principle capable of being appealed to by all parties concerned, some principle from which we can reason in common.*[11]

This idea of reasoning in common with others is for Rawls one of the two "basic" aspects of what it means to be reasonable; the second is "the willingness to recognize the burdens of judgment and to accept their consequences for the use of public reason in directing the legitimate exercise of political power in a constitutional regime."[12] For present purposes, however, the first aspect is clearly the more important of the two. So understood, "the reasonable" is at the heart of his conception of society as a system of fair cooperation over time. Reasonable persons thus are those who

---

[6] *Id.* at 14–15.

[7] *Id.* at 10–15.

[8] *Id.*; *see also* JOHN RAWLS, JUSTICE AS FAIRNESS: A RESTATEMENT (Erin Kelly ed., 2001).

[9] See Morris, *supra* note 3, at 57 & n.11, where Morris made the following observation about Rawls's statement that the theory of justice may be the most important part of rational-choice theory: "It is clear from his most recent writings, if not from some of the elements of *A Theory of Justice*, that Rawls does not really endorse the view of moral theory expressed by his remark."

[10] JOHN RAWLS, POLITICAL LIBERALISM 48–54 (1993).

[11] W.M. Sibley, *The Rational Versus the Reasonable*, 62 PHIL. REV. 554, 557 (1953) (emphasis added to last sentence).

[12] RAWLS, *supra* note 10, at 54.

> are ready to propose principles and standards as fair terms of cooperation and to abide by them willingly, given the assurance that others will likewise do so. Those norms they view as reasonable for everyone to accept and therefore as justifiable to them; and they are ready to discuss the fair terms that others propose.[13]

To be reasonable, in short, is to be committed to fair cooperation. This is a commitment neither to altruism nor to the rational pursuit of one's interests, but to the middle ground of reciprocity:

> Reasonable persons . . . are not moved by the general good as such but desire for its own sake a social world in which they, as free and equal, can cooperate with others on terms all can accept. They insist that reciprocity should hold within that world so that each benefits along with others.[14]

For Rawls, then, justice as fairness becomes a part of the theory of reasonable, rather than rational, choice. Many of those who are taken with the prospects of social contract theory have followed him down this path, but others have resisted. The result is that there are now two contending groups of contract theorists: one known as the *contractarians* and another called the *contractualists*.[15] There are variations and intramural disputes within each group, but the major differences between the two are all that need concern us here.

One way to mark the difference between these groups is to say that the contractarians take a Hobbesian approach while the contractualists are more likely to follow the lead of Kant or Rousseau.[16] Like Thomas Hobbes, the contractarians begin by assuming that the parties to the social contract are rational agents seeking to advance their interests; whereas Hobbes argued that the foundation of political authority is in the social contract, however, the contractarians propose to derive the fundamental terms of social life—that is, *morality*—from the contract. Hence the title of what is probably the most influential work in this category, David Gauthier's *Morals by Agreement*.[17] For their part, contractualists tend to divide into two overlapping categories: those who concentrate, as Rawls does, on devising a theory of social or political justice, and those who follow Thomas Scanlon in arguing that contractualism provides the underpinnings of an adequate

---

[13] *Id.* at 49.

[14] *Id.* at 50.

[15] For helpful, concise accounts of these two positions, see Ann Cudd, *Contractarianism*, STAN. ENCYCLOPEDIA PHIL. (Apr. 4, 2007), http://plato.stanford.edu/entries/contractarianism, and Elizabeth Ashford & Tim Mulgan, *Contractualism*, STAN. ENCYCLOPEDIA PHIL. (Aug. 30, 2007), http://plato.stanford.edu/entries/contractualism.

[16] *See* Ashford & Mulgan, *supra* note 15.

[17] DAVID GAUTHIER, MORALS BY AGREEMENT (1986).

theory of morality as a whole.[18] Whether their concern is primarily political or moral, though, their starting point is the reasonable individual who, already engaged in social cooperation, wants to find and live according to fair or just principles in reciprocity with others.

Whether either of these philosophical approaches to moral and political problems is satisfactory is a matter of considerable debate. To their critics, each of them seems to suffer from a crippling defect. In the case of the contractarians, the question is whether rational individuals seeking to maximize their interests would indeed generate terms of social cooperation that are recognizable as morality.[19] It may be rational, as Gauthier and others contend, for individuals to constrain their attempts to maximize their interests in order to enjoy the benefits of social cooperation—and to avoid the horrors of something like a Hobbesian state of nature—but it appears to be even more rational to endorse a rule on the order of, "Constrain maximization when one must, but maximize one's interest by being a free rider who takes advantage of the cooperative efforts of others whenever possible." Such a rule, of course, is more likely to subvert than to support social cooperation and is quite unlikely to provide the foundations of morality.

Contractualists do not face this criticism, for they begin with reasonable individuals who are already committed to doing their part in a fair system of social cooperation. In their case, though, the worry is that their starting point simply assumes too much. The contractarian at least tries to provide an answer to the old question: Why be moral? The contractualist, however, dodges this question by assuming that reasonable people will be moved by the demands of morality or justice. For the contractualist, the important question is how to work out what those demands require of us.[20]

These criticisms of the two contract-based theories will play a part in the discussion to follow. At this point, though, it is best to turn from the broader concerns of moral and political philosophy to the particular question of legal philosophy of special interest here: How, if at all, can we justify the punishment of criminals? If either contractarianism or contractualism can provide the basis for a compelling answer to this question, it will give us a reason to put our general worries about the theory on hold while we pursue its implications for other long-standing problems. But does either theory have a compelling answer?

---

[18] T.M. Scanlon, *Contractualism and Utilitarianism*, *in* UTILITARIANISM AND BEYOND 103, 103–04 (Amartya Sen & Bernard Williams eds., 1982) ("Despite the wide discussion which [*A Theory of Justice*] has received, however, I think that the appeal of contractualism as a foundational view has been underrated. In particular, it has not been sufficiently appreciated that contractualism offers a particularly plausible account of moral motivation."); *see also* T.M. SCANLON, WHAT WE OWE TO EACH OTHER (1998).

[19] *See, e.g.*, Cudd, *supra* note 15; Frank Lovett, *Justice, Theories of*, *in* 2 ENCYCLOPEDIA OF POLITICAL THEORY 733, 738 (Mark Bevir ed., 2010).

[20] *See, e.g.*, Frank Lovett, *Can Justice Be Based on Consent?*, 12 J. POL. PHIL. 79 (2004).

III. CONTRACTARIANISM AND PUNISHMENT

Neither Morris nor Finkelstein explicitly identifies his or her position as contractarian in the sense sketched above. Morris published his essay in 1991, before the distinction between contractarianism and contractualism had become conventional among moral and political philosophers, and Finkelstein is not concerned to place herself into one of the two camps. Yet neither of them adheres to the reasonability standard that Brettschneider takes to be the hallmark of the contractualist approach.[21] When applied to punishment, this standard, or "'contractualist' test," requires us to determine whether a particular instance of coercion is justified by asking, "Given a motivation to reach universal agreement, could citizens who view themselves as free and equal *reasonably* reject such an instance of coercion?"[22] For this and other reasons that should become evident below, it is appropriate to locate Morris and Finkelstein on the contractarian side of the divide.

A. *Morris on Moral Standing*

Like many others who have written about the justification of punishment, Morris begins his essay by asking, "By what authority do we punish? What permits us to deprive people of their liberty or possessions for some wrong that they have committed?"[23] His answer takes the form of a "forfeiture" account: "criminal acts alter the moral status of wrongdoers, [so] that such acts affect their moral rights and lead to their forfeiture."[24] Rather than defend this claim "merely at an intuitive level," however, Morris aims to provide "a theoretical account of the relationship between action and intention and the moral status of agents."[25]

In developing this theoretical account, Morris associates himself with "a long western tradition" that takes justice to consist of "principles, rules, and norms that ideally serve to advance the interests and aims of all in certain situations. This tradition is dubbed 'contractarian' as it often understands the terms of justice to be the outcome of a hypothetical bargain or 'social contract.'"[26] But it may be "less misleading," he says, "to think of contemporary representatives of this tradition as offering a 'rational choice' conception of morality."[27] Whatever we call it, following this tradition, with its attendant conception of morality, will lead us to conclude that "in the circumstances of justice all humans capable and willing to impose constraints on their behaviour toward others have full moral standing"; to have this standing is to enjoy a set of moral rights—"the most important of which,

---

[21] *See* Brettschneider, *supra* note 4, at 176–77.
[22] *Id.* at 177–78.
[23] Morris, *supra* note 3, at 53.
[24] *Id.* at 54.
[25] *Id.*
[26] *Id.* at 57.
[27] *Id.*

we may assume, are those to life, liberty, and property"—and to "lose some such rights is to lose some of one's moral standing."[28]

According to Morris, then, the terms of morality in general and justice in particular are those that rational individuals would choose to impose on themselves in a hypothetical bargaining situation. The parties to this bargain, or contract, will have full and equal moral standing—and thus rights to life, liberty, and property—so long as they abide by its terms. Individuals will be tempted, though, to ignore or violate the terms of the contract when doing so appears to advance their interests or satisfy their desires; if they prove willing to act on this temptation, they forfeit some or all of their moral standing, and if their acts are serious enough to be deemed crimes, they make themselves liable to condign punishment. As Morris says, "Individuals unwilling to respect such constraints [as the contract imposes] will lack, or lose, full moral standing. One expression of such unwillingness, we may suppose, is the intentional commitment of certain crimes. Thus, certain criminal acts will deprive their performers of some part of their moral standing."[29]

Taken to the extreme, Morris notes, his version of a contractarian theory of punishment may be used to justify capital punishment or torture, although he himself does not endorse such practices.[30] The point is that

> we do not have to give standard *moral* justifications for executing contract killers, war criminals, tyrants, or terrorists because so killing them would neither be a violation nor an overriding of their moral rights to life or liberty. Rather, they no longer have, or never had, such moral rights. Thus we merely need sufficient reason to execute them.[31]

Morris's contractarian theory is in many ways quite attractive as a justification of punishment, but most especially so because of its ability to accommodate what seem to be core intuitions about punishment. One of these intuitions rests on the distinction between *mala in se* and *mala prohibita* offenses—a distinction that Morris does not even mention in his essay. Here, the intuition is that there are some kinds of actions that warrant punishment as failures of cooperation or respect for others even though, as *mala prohibita*, they are not the kinds of actions that are paradigmatic crimes, such as murder, rape, robbery, and assault.[32] Morris's theory can sustain this intuition by explaining that individuals who fail to play their parts in the cooperative arrangements that follow

---

[28] *Id.* at 62.

[29] *Id.* at 65.

[30] *Id.* at 74 n.44.

[31] *Id.* at 74.

[32] For further thoughts on the relationship between *mala prohibita* and paradigmatic crimes, see S.E. Marshall & R.A. Duff, *Criminalization and Sharing Wrongs*, 11 CANADIAN J.L. & JURISPRUDENCE 7, 19–22 (1998), and Richard Dagger, *Republicanism and Crime*, *in* LEGAL REPUBLICANISM: NATIONAL AND INTERNATIONAL PERSPECTIVES 147, 158–66 (Samantha Besson & José Luis Martí eds., 2009).

from the hypothetical contract surrender some of their moral status and thereby justify their punishment because their failure to cooperate constitutes a *malum prohibitum*, such as tax evasion or driving on the wrong side of the road. The same is true, of course, and explicitly so, of those who commit *mala in se* offenses—not only the "contract killers, war criminals, tyrants, and terrorists" of whom Morris speaks but also those we call "common criminals." Here again, the intuition is that such people forfeit some or all of their moral status, with the attendant rights to life, liberty, and property, when they prove themselves unworthy by their failure to respect the moral status of others.

The question for Morris, however, is whether his justification of punishment is attractive because it provides a philosophical foundation for these intuitions or because it trades on their appeal. In other words, does the contract or hypothetical bargaining situation do any real work, or is it appealing simply because it seems to correspond with our preexisting intuitions about when punishment is justified? I believe that the latter is the correct answer.

To be sure, one problem for Morris is the general objection to contractarianism mentioned earlier—that is, that rational agents in the hypothetical bargaining situation will not so constrain themselves as to foreclose the option of free riding, should the opportunity present itself. To this complaint Morris could respond that "rational choice or contractarian moral theory understands rational humans, capable *and willing* to impose moral constraints on their conduct toward others, as moral subjects and direct moral objects"—that is, as having moral standing.[33] This is to say that only those who commit themselves to complying with the agreed upon principles and rules would have full moral standing; free riders would not. Whether this is a satisfactory response, however, is not obvious. Indeed, this account of moral standing seems to be a matter of stipulation or perhaps an acknowledgment of common intuitions rather than a product of contractarian theory.

In this regard it is useful to consider what Morris says about natural rights theories, from which he distinguishes his own approach:

> It is often claimed that we possess certain moral rights by virtue of the sort of creature that we are, that is, *by virtue of our nature*. This claim is explicit in virtually all accounts of *natural* rights . . . . Now rights that are acquired by virtue of the possession of certain (natural) attributes (e.g., rationality, self-consciousness) cannot be lost except by the loss of those very attributes. On such an account criminals retain their moral rights, as long as they do not lose their rationality or whatever natural attributes generate moral rights. Moral standing is usually thus understood in terms of possession of certain attributes.
>
> While possession of certain natural attributes (e.g., rationality) is *necessary* for moral standing on the account that I am offering, it is not

[33] Morris, *supra* note 3, at 60 (emphasis added).

> *sufficient.* The reason for this is that justice is to be understood as a mutually beneficial convention which constrains rational agents in situations where individually rational action leads to disadvantageous outcomes. Compliance with the requirements of justice is the condition for protection by those requirements.[34]

So stated, Morris's theory does appear to have the advantage on the natural rights approach. It is remarkable, though, that Morris does not mention one of the preeminent natural rights theorists, John Locke, in this regard; nor elsewhere in his essay. The omission is remarkable because Locke's position on the moral standing of criminals seems much the same as Morris's. As Locke says in his *Second Treatise* (§ 11),

> [E]very man, in the state of nature, has a power to kill a murderer, both *to deter* others from doing the like injury . . . and also to secure men from the attempts of a criminal, who having renounced reason, the common rule and measure God hath given to mankind, hath, by the unjust violence and slaughter he hath committed upon one, declared war against all mankind, and therefore may be destroyed as a *lion* or a *tyger*, one of those wild savage beasts, with whom men can have no society nor security.[35]

For Locke as for Morris, then, the criminal who commits "unjust violence and slaughter" loses moral standing; both thus advance a forfeiture account of rights or moral standing. In Locke's colorful language, in fact, the criminal seems no longer to be fully human but to have degraded himself to the status of the lion or tiger "with whom men can have no society nor security." Locke's account differs from Morris's, however, in that Locke makes no appeal to a contract or convention to justify the judgment that the criminal has lost moral standing. Locke is a social contract theorist, to be sure, but in this case, as the quotation above indicates, everything takes place in the state of nature. For Locke, criminals lose some or all of their natural rights, and thereby justify others in punishing them, even before anyone enters into the social contract. This being so, there is no need to invoke a contractarian theory in order to explain how punishment is justified as the proper response to those who have lost moral standing.

That conclusion only follows, of course, if Locke's argument is sound, and Morris might well deny that it is. He might claim, for instance, that Locke is one of those natural rights theorists who provide a necessary but not sufficient reason for holding that criminals forfeit their moral standing. After all, Locke himself said that the criminal has "renounced reason," which is to say that the criminal has

---

[34] *Id.* at 65–66 (footnotes omitted).

[35] JOHN LOCKE, SECOND TREATISE OF GOVERNMENT § 11, at 11 (C.B. Macpherson ed., Hackett Publishing Co. 1980) (1690).

lost the natural attribute of rationality that underpins his or her moral standing, but that is not sufficient to show that the criminal has acted unjustly and therefore deserves punishment. Morris's general objection to natural rights accounts of punishment applies to Locke, then, as it does to other natural rights theorists.

There are two problems, however, with this line of response. One is that the criminal in Locke's state of nature who has "renounced reason" does not seem to have lost the natural attribute of *rationality*. Indeed, the criminal may be acting quite rationally, in the instrumental sense of the word, when he kills someone who stands in the way of achieving what he wants. We cannot be sure here, and there is a risk of reading Locke anachronistically, but there is a hint of the distinction between the rational and the reasonable, as the passage quoted above indicates, in his account of crime and punishment in the state of nature. If so, then Locke's claim that the criminal has renounced *reason* looks less like the loss of a natural attribute than sheer unwillingness to give due consideration to the rights and interests of others. It looks, that is, like the kind of *injustice* that provides a sufficient justification, on Morris's account, for punishment. If so, then it is injustice that is not explained in contractarian terms.

How, then, is the injustice explained? For Locke, the explanation is that the criminal has violated the law of nature, which forbids us—"unless it be to do justice on an offender"—to "take away, or impair the life, or what tends to the preservation of the life, the liberty, health, limb, or goods of another."[36] No contract or convention is necessary. The second problem that Locke's argument presents to Morris, then, is that Morris must demonstrate that the hypothetical contract is necessary to the justification of punishment. If contractarianism is sufficient, in other words, but no more sufficient than another theory (or theories) that can do the same thing, then why should we prefer it to the other theory (or theories)? Why not simply follow Locke and others who argue in terms of natural law, or—as I shall suggest later—turn to a theory of punishment based on the principle of fair play?

The standard answer to this kind of challenge is to say that the theory one wants to defend not only can do what its rivals do but that it can also do something more or better than they do. That is the challenge that Morris and those who follow a similar version of contractarianism must meet. They may also, however, need to respond to a different challenge that may be raised from within the contractarian ranks as easily as from without.

B. *Finkelstein on Consenting to Punishment*

This second challenge arises from those who do not agree that wrongdoers forfeit some or all of their moral standing when they commit crimes. Morris seems to anticipate this challenge from advocates of natural rights who will insist, as he says, that "criminals retain their moral rights, as long as they do not lose their

---

[36] *Id.* at 9.

rationality or whatever natural attributes generate moral rights."[37] His response seems to be that those who take this position must regard the punishment as a matter of *overriding* or *justifiably infringing* the criminals' rights. The latter view he finds implausible, and the former—that criminals retain their rights even when some of them are overridden—seems not to apply when indefeasible rights, such as the right to life, are in question.[38]

What Morris does not anticipate, however, is that this challenge may arise from a fellow contractarian, as it does in Claire Finkelstein's *A Contractarian Approach to Punishment*.[39] Finkelstein cites Morris's *Punishment and Loss of Moral Standing* as an example of "the usual approach to punishment in the contractarian tradition"—a tradition that "treats violators of the social contract as permanently expelled from the contractual relationship that holds among members of society. The tradition thus denies that punishment is governed by the terms of the contract itself, and treats it as governed by norms that lie outside the contract."[40] Against this view, Finkelstein maintains that it "would be wrong to treat defection as though it were noncooperation at the outset."[41] She offers three reasons in support of this conclusion,[42] two of which seem to me misplaced, and one of which is telling, although not perhaps in the way that Finkelstein intends.

It is not my purpose to arbitrate disputes among contractarians, but it does seem that there are issues of extramural significance here. In the broadest terms, there is the question of whether allowing for the loss or forfeiture of moral standing, as Morris does, is an advantage or a defect in a theory of punishment. Contractarian or not, those who believe that allowing for forfeiture is an advantage, as I do, will want to know whether Finkelstein's criticism of that approach is sound. I doubt that it is. To begin with, I do not see that Morris's version of contractarianism entails that "violators of the social contract" be treated "as permanently expelled from the contractual relationship that holds among members of society."[43] Morris holds that "wrongdoers lose *some* of their rights and *some* of their moral standing, and that some wrongdoers lose all of their rights (or never possessed the full set) and retain at most . . . partial moral standing."[44] What follows, then, is that *some* of those who violate the social contract, but by no means *all of them*, are in a sense to be expelled from the contractual relationship. Those who are not expelled—and I suspect that Morris would include the great

---

[37] Morris, *supra* note 3, at 65.

[38] *Id.* at 56; *see also id.* at 54 (discussing the implausibility of justified infringement of rights); *id.* at 74 (noting the indefeasibility of a right to life).

[39] Finkelstein, *supra* note 3. Professor Finkelstein's contribution to the present symposium provides a helpful elaboration of the theory sketched in her earlier essay, but I believe it leaves the theory vulnerable to the criticisms I advance below. *See* Claire Finkelstein, *Punishment as Contract*, 8 OHIO ST. J. CRIM. L. 319 (2011).

[40] *Id.* at 217–18.

[41] *Id.* at 218.

[42] *Id.*

[43] *Id.* at 217–18.

[44] Morris, *supra* note 3, at 62–63 (emphasis added).

majority of criminals in this category—will remain within the contract while suffering some loss of their rights, such as the right to be at liberty, for some period of time. Nor do I see anything in Morris's argument that would prevent him from endorsing forms of punishment that aim at restoring criminals to full moral standing and full membership in the community once their punishment is complete.

That is why I think that Finkelstein's first two reasons for rejecting the "usual" contractarian position on punishment are misplaced. The first reason she offers is that "defections can be large or small, and it may be that it is still advantageous to cooperate with those responsible for small defections."[45] She is surely right on this point, but Morris and others who follow the "usual" contractarian tradition have no reason to disagree. Morris can approve the loss of a litterer's right to some of her property in the form of a fine, for example, or the loss of some of her right to liberty in the form of community service, without banning her altogether from social or legal cooperation.[46] Finkelstein's second reason for rejecting the "usual" contractarian position is also correct but misplaced. As she puts the point, "it is not possible to address the problem of noncooperation at the outset in any way other than refusing to contract. But defectors are themselves subject to the terms of an antecedent agreement, and can therefore be dealt with contractually."[47] In this case the criticism is misplaced because Morris does not couch his argument in terms of cooperators and defectors. Talking about contracts may often lead to the use of these other terms, but that is not where it leads Morris. On his account, the people who lose some or all of their rights and open themselves to punishment are not "defectors" but "wrongdoers"; they are those who violate or demonstrate "an unwillingness to abide by the constraints of justice."[48] What counts is not defection from a contract that one has accepted but failure to live within the constraints of justice that rational agents would choose in a hypothetical bargaining situation. There is thus no opportunity at the outset to choose either to cooperate or to refuse cooperation and thus no reason to distinguish cooperators from defectors.

Even if we dismiss these two objections, however, the third remains, and it may be sufficient support for Finkelstein's rejection of the "usual" contractarian position—at least if we substitute a term such as "wrongdoer" or "unjust person" for "defector." Her objection here is straightforward:

> [I]t simply seems wrong to think of a defector as beyond the bounds of all social interaction . . . . Even the most heinous violations ought not to deprive their perpetrators of basic dignitary rights, such as the right to be

---

[45] Finkelstein, *supra* note 3, at 218.

[46] See Morris, *supra* note 3, at 78–79, for his comments on "outlawry."

[47] Finkelstein, *supra* note 3, at 218.

[48] Morris, *supra* note 3, at 69.

> free from torture, the right to speak in one's own defense, and the right to minimal bodily dignity and comfort.[49]

Again, we should note that Morris does not hold that all wrongdoers are "beyond the bounds of all social interaction." That conceded, however, there does seem to be a fundamental difference in some cases between Morris and Finkelstein. This is a telling point, in my view, because there seems to be no basis for settling the disagreement within the contractarian framework. Morris can say that rational agents in the hypothetical contractual situation would agree that the most heinous violators of morality and justice have lost (or never had) full moral standing, but Finkelstein apparently will respond that it would be wrong of them to do so—that this is something to which they simply could not agree. Who is right? The answer apparently depends on what one thinks rational agents would agree to, and that, in turn, seems to depend on what one thinks, without resort to contractarian reasoning, about the possibility of losing or forfeiting moral rights. All of which is to say that contractarian reasoning does no real work here. What matters are one's intuitions or convictions about rights and punishment, for those will shape the results that emerge from the hypothetical contract.

Finkelstein's third objection is thus, as I have said, a telling one, but it is also a double-edged sword. It is effective, that is, against Morris and the "usual" contractarian position on punishment, but it is also effective against her contractarian account of punishment. When an argument rests on the claim that "it simply seems wrong," we have reason to believe that the moral intuition or conviction is doing the heavy lifting rather than the contractarian reasoning. To be fair, Finkelstein does go on to say that "the conditions under which human beings may permissibly inflict sanctions for noncooperation on members of their own kind should be thought of as governed by *an antecedent agreement such humans make to enforce the terms of cooperative interaction*."[50] But what would govern the terms of *that* agreement? Would the parties involved in it be allowed to deprive heinous wrongdoers of the basic dignitary rights? Or would Finkelstein say, again, that it would be wrong for them to do so? At some point, if Finkelstein is right, that will have to be the fundamental judgment: "it simply seems wrong" to do so.

Perhaps, though, we should give the benefit of the doubt to Finkelstein's attempt to develop a contractarian account of punishment that differs significantly from the "usual" position. She certainly differs from Morris in the weight she gives to *consent* and *benefit* in her justification of punishment. Perhaps this difference will provide the basis for a justification of punishment in which the idea of a contract will truly be significant.

The importance of consent to Finkelstein's account of punishment is evident in the fact that she refers to her position as both "a contractarian approach" and "a

---

[49] Finkelstein, *supra* note 3, at 218.

[50] *Id.* (emphasis added).

consensual theory of punishment."[51] Her argument is that "no treatment of another human being as harsh as that which standard forms of punishment for serious crimes involve can be permissible if it is truly involuntarily imposed. For this reason, only a consensual theory of punishment holds out hope for a true justification for the institution."[52] She quickly adds, though, that consent alone is not sufficient to provide this justification; consent must be "coupled with the fact that the agent receives a benefit under the institution to which he consents."[53] This argument will no doubt seem as counterintuitive to many as Rousseau's claim that the murderer consents to his execution, but Finkelstein seems to think that it is at least as counterintuitive to believe that we can justifiably impose harsh treatment on someone who has not somehow agreed to it. Her defense of her position, which appeals to Rawls and Hobbes rather than Rousseau, proceeds as follows.

First, she asks us to begin by assuming that society is, "to use Rawls' phrase, 'a cooperative venture for mutual advantage.'"[54] We then should ask, what kinds of institutions for the basic structure of society would rational agents agree to accept were they charged with forging a social contract? With regard to punishment, the question would be: "would each rational agent involved in selecting the basic institutions of society regard it as advantageous to include punishment among those to which she gives her assent?"[55] The answer turns out to be yes because life in a society without the institution of punishment would prove to be intolerable. Each rational agent will recognize that the pain of punishment may fall upon her at some point, but she will also recognize that she will be better off in a society that practices punishment than in one that does not. The institution of punishment thus passes "the benefit test" and warrants the consent of rational agents:

> Thus each member of society must project himself into the position of someone who has violated the conditions of the more basic, substantive social contract, and ask himself whether, if he were to be punished for such violations, he would still fare better than he would had he never agreed to live under threat of punishment in the first place. For many sanctions the benefit test will be satisfied. . . . [F]or most penalties, and most societies, even an offender who must suffer punitive sanctions will fare better under a punishment agreement than he would in the absence of all social enforcement.[56]

As Finkelstein's use of "many" and "most" suggests, her consent-plus-benefit theory is not likely to approve every form or extent of punishment. She does say

---

[51] *Id.* at 207; *id.* at 208.
[52] *Id.* at 208.
[53] *Id.*
[54] *Id.* at 214 (quoting RAWLS, *supra* note 5, at 4).
[55] *Id.*
[56] *Id.* at 215–16.

that punishment is a valuable institution because of its deterrent effect, and if it succeeds in deterring crime against a person until he or she commits murder at, say, the age of twenty-five, then even the death sentence or life in prison will not obviously outweigh the benefits one received from the existence of punishment in those twenty-five years of life, for those are probably twenty-five more than he or she would have enjoyed in a Hobbesian state of nature.[57] But Finkelstein also notes that the consent-plus-benefit theory will not condone a harsher penalty when a less severe one will have almost the same deterrent effect. As she says, "If the death penalty has only very modest additional deterrent efficacy over life in prison without parole, it is unlikely to be incorporated into the punishment agreement, as its detriment for the person suffering it is vastly greater than the nearest available alternative penalties."[58] Consent-plus-benefit thus may prove to have the attractive quality of enabling us to discriminate between forms and degrees of punishment that are warranted and those that are not.

Notwithstanding that attractive quality, however, Finkelstein's account seems doubly flawed. First, it seems wrong to insist that the punished person must receive more overall benefit than suffering from the institution of punishment if his or her punishment is to be justified. Finkelstein herself indicates that her theory faces a problem here when she considers the case of "a very young offender" who has not lived long enough to reap much benefit from the deterrent effects of punishment and "now could reap no further benefits from such rules if put to death or imprisoned for the rest of his life."[59] Accepting such a possibility, though, runs counter to the intent of her consent-plus-benefit theory. That is, it would seem that the theory should prohibit punishing anyone in a way that would leave him with more losses than gains in the benefit column at the end of his life. That is why Finkelstein suggests, as we have seen, that the death penalty is unlikely to be approved by the parties to the contract. Should the same reasoning apply to life in prison without possibility of parole, at least for criminals who are yet to reach a certain age? Perhaps, but Finkelstein does not draw this conclusion, which leaves her in the difficult position of approving a form of punishment that seems at odds with her theory. If she does go on to endorse this conclusion, though, she will still face difficulties in this regard. For it is simply not clear that it is wrong to impose on someone a punishment that would outweigh the benefits he received from the existence of punishment as an institution. Sympathies and intuitions may differ on this point, but surely many people would maintain that heinous offenders—even very young ones—deserve to suffer for their crimes no matter how little benefit they received from the deterrent effect of punishment in their own lives. The question of whether the offender has benefited, on this view, is simply beside the point.

---

[57] *See id.* at 216.
[58] *Id.*
[59] *Id.*

Finkelstein can respond, of course, that the whole point of her theory is to deal with objections such as this by appeal to contractarian reasoning. Those who maintain that the degree of the offender's benefit is beside the point would have to change their minds, in other words, were they in the position of rational agents who must select the basic institutions of society. But here we see the second flaw in Finkelstein's consent-plus-benefit approach to punishment. Finkelstein maintains that rational agents would only agree to institutions and rules that leave them with a net gain of benefits over losses. But that conclusion requires, as we have seen, that

> each member of the society must project himself into the position of someone who has violated the conditions of the more basic, substantive social contract, and ask himself whether, if he were to be punished for such violations, he would still fare better than he would have had he never agreed to live under threat of punishment in the first place.[60]

That is to say, in effect, that the institutions of punishment must be designed from the criminals' point of view. But why should that be the case? Why not have the rational agents project themselves into the position of those who are victims rather than criminals? What would the outcome be in that case? For that matter, shouldn't the rational agents project themselves into the position of (potential) criminals and (potential) victims, then define the allowable forms and extents of punishment from that combination of perspectives? What then would the outcome be? Would it be guaranteed to be consistent with Finkelstein's desire to show that punishment must be, on balance, in the interests of even the one who is punished?

To pursue these questions further would take us into complicated matters of decision theory like those surrounding Rawls's adoption of the maximin strategy to justify his "difference principle" in *A Theory of Justice*[61]—matters that are well beyond the scope of this essay. But the key point, I take it, is established. Contractarian reasoning by itself will not generate the conclusions that Finkelstein wants. It will not show us that it is simply wrong to treat some criminals as if they have lost or forfeited their moral standing, and it will not show us that criminals must receive not merely *a benefit* from the practice of punishment but a *net benefit*.

In this respect, Finkelstein's account of punishment is in the same position as that of her contractarian counterpart, Christopher Morris, for neither has demonstrated that contractarian reasoning itself generates the conclusions he or she wants. But what of their contractualist rivals? Do they have an account of punishment that truly derives from something like a social contract?

---

[60] *Id.* at 215–16.

[61] *See* RAWLS, *supra* note 5, at 152–57. See also Finkelstein, *supra* note 3, at 215, for her Rawlsian defense of her "no-gambling requirement."

IV. CONTRACTUALISM AND PUNISHMENT

Contractualists have had much to say about morality, justice, and political legitimacy but little about punishment or other responses to crime. The significant exception is Corey Brettschneider's recent essay, *The Rights of the Guilty: Punishment and Political Legitimacy*.[62] As the subtitle indicates, Brettschneider's aim is to link the justification of punishment to the legitimacy of the state or polity that carries out the punishment. Contractualism provides the standard of political legitimacy, he holds, in the form of "Rawls's 'liberal principle of legitimacy,'" which Brettschneider quotes as follows: "'our exercise of political power is fully proper only when it is exercised in accordance with a constitution the essentials of which all citizens as free and equal may reasonably be expected to endorse in light of the principles and ideals acceptable to their common human reason.'"[63] Central to this principle, on Brettschneider's account, is the belief that political coercion must be justified "to those who are guilty of crimes."[64]

Brettschneider's account of punishment resembles Finkelstein's in this respect, but he reaches the conclusion by appealing not to rational agents, as she does, but to *reasonable citizens*.[65] The starting point for contractualism, in other words, is not the rational agent concerned to create rules to govern social interaction that will most fully advance his or her interests; it is, instead, the reasonable citizen who wants to do his or her part in maintaining a fair system of social cooperation over time. The individuals in the contractualist thought experiment are agents, to be sure, but they are agents who think of themselves as *citizens* rather than *persons* or *actual individuals*.[66] Their focus, then, is on what they share with other members of the polity, whom they conceive of as free and equal citizens engaged in a reciprocal relationship with one another. When proposing or assessing proposals for principles, rules, or laws to govern the polity, the reasonable citizen thus will ask whether the proposal in question is one that citizens can reasonably accept as in keeping with everyone's status as a free and equal citizen.[67] If the answer is no—if some citizens can reasonably reject the proposal—then it cannot be enacted or approved.[68]

Applying contractualist reasoning to punishment leads Brettschneider to frame his approach in the following way:

---

[62] Brettschneider, *supra* note 4.

[63] *Id.* at 176 (quoting RAWLS, *supra* note 10, at 137).

[64] *Id.*

[65] *Id.* at 177.

[66] *Id.* at 178.

[67] *See id.* at 177–78.

[68] There is a debate among contractualists as to whether the standard is reasonable acceptance or reasonable rejection, but that need not concern us here. Brettschneider does not join the debate, and he appeals to both acceptance and rejection, as subsequent quotations will indicate.

> The issue for contractualist thinkers is not whether criminals, given their empirical disposition (or lack thereof) toward reasonableness, would actually accept a punishment or not. Rather, the issue is whether a particular criminal who has committed a particular act could *reasonably* accept a given punishment. In other words, contractualist justification is concerned with punishment addressed to the criminal *qua* citizen and whether those who have committed crimes could reasonably accept such punishments. The goal is not to legitimize only those punishments that criminals would actually accept but rather to assess which punishments a criminal might reasonably accept were she motivated to find universal agreement about how to balance her interests with the interests of others.[69]

Brettschneider refers to neither Morris nor Finkelstein in his essay, but his contractualist approach to punishment leads to conclusions that differ in significant ways from theirs—just as their conclusions differ in interesting ways from each other. Like Morris, Brettschneider concludes that convicted criminals must forfeit some of their rights as citizens, but he agrees with Finkelstein in holding that "it is too extreme to claim that they [i.e., convicted murderers] must forfeit all of their rights associated with political legitimacy."[70] Morris allows for a complete forfeiture, as we have seen, when he argues that "contract killers, war criminals, tyrants, and terrorists" may lose their moral standing and rights.[71] There is also the further difference that Morris is concerned with *moral* rights and standing, whereas Brettschneider's concern is for the criminal's rights and standing as a *citizen*—with rights "associated with political legitimacy." For Brettschneider, in fact, it is important to develop a distinctly *political* theory of punishment that will not only address the moral issue of what a criminal deserves but also the question of "what punishments the state can rightly mete out in a way consistent with our general ambition to distinguish legitimate state action from brute power."[72]

Brettschneider's contractualist reliance on reasonable citizens rather than rational agents also leads to a significant difference from Finkelstein with regard to how the criminal's point of view is to be taken into account. They both differ from Morris in giving some consideration to the criminal's perspective, but Brettschneider does not require that the application of punishment must meet Finkelstein's "benefit test"—that is, that the form or extent of punishment must not lead to a net loss of benefit to the criminal. For Brettschneider, the criminal's perspective is to be considered, but it is the perspective of the "criminal *qua* citizen" who is "motivated to find universal agreement about how to balance her interests with the interests of others."[73] Whether contractualist reasoning will

---

[69] Brettschneider, *supra* note 4, at 179.
[70] *Id.* at 180.
[71] Morris, *supra* note 3, at 74.
[72] Brettschneider, *supra* note 4, at 194; *see also id.* at 175.
[73] *Id.* at 179.

produce an outcome significantly different from Finkelstein's contractarian approach is not altogether clear, but Brettschneider at least gives consideration to other perspectives than the criminals', as is evident in the following scenario:

> We can imagine that at the moment of sentencing, the judge agrees to hear the defendant make arguments about his or her reasonable interests. However, those arguments should acknowledge that the interests of his or her fellow citizens also have weight in determining the answer to this question. The judge acts as a representative of the community by recognizing the reasonable interests of the victims and potential victims and balancing them with those of the defendant when formulating the sentence. All must be included in determining which sentence is compatible with mutual justification. This process models reciprocity because the defendant gives reasons designed to be reasonably acceptable to the judge (as a representative of the community). Likewise, the judge engages in a form of justification aimed at being reasonably acceptable to the criminal.[74]

What exactly will be *reasonably* acceptable to the criminal will vary, of course, from case to case, but it seems clear that this standard will allow for punishments that Finkelstein's "benefit principle" would rule out. What the reasonable criminal *qua* citizen in the above scenario must accept is likely to be considerably harsher than what rational agents, reasoning from the criminal's point of view, would agree to in a hypothetical bargaining situation.

That is not to say, though, that Brettschneider believes that contractualist reasoning will justify a regime of unrelenting punishment. On the contrary, he indicates that reasoning about punishment in this way will lead to significant limits on the forms and extent of punishment. In particular, he advances four claims. First, punishments that involve "wanton violence" or "deliberately inflict pain without regard for both the interests of the criminal and the interests of society are reasonably rejected."[75] Second, in keeping with the idea that criminals do not forfeit all of their rights, convicted criminals can reasonably retain some of their rights to freedom of speech. These will need to be balanced against the need to maintain security in prisons and jails, but "prison policy should . . . infringe as little as possible on the legitimate exchange of ideas, especially in [prisoners'] communication with the outside world."[76] Third, with regard to laws that deprive felons of the right to vote, Brettschneider argues that "those who have served their time could reasonably reject attempts to deny them the franchise even after they are released from prison"; otherwise, "their punishment would create a group of second-class citizens who are not treated as free and equal."[77] It may even be

---

[74] *Id.* at 186–87 (footnote omitted).
[75] *Id.* at 188.
[76] *Id.* at 189.
[77] *Id.*

reasonable to allow prisoners to vote in some elections while they are serving their sentences.[78] Finally, capital punishment is "entirely inconsistent" with the "ideals of free and equal citizenship."[79]

Assuming that these four claims are not so obviously wrong as to make further discussion pointless, two questions arise with regard to them. The first is whether contractualist reasoning necessarily leads to these conclusions, and the second is whether one can only reach these conclusions by means of contractualist reasoning. Brettschneider's answer to the first question seems to be that contractualist reasoning is fully conclusive in some cases, such as capital punishment, but only partially so in others. In the case of the claims to freedom of speech and the right to vote, the point seems to be that prisoners must retain their democratic rights so far as possible, leaving contractualists to argue among themselves about just how far it is reasonable to go in this respect. In the case of capital punishment, however, Brettschneider concludes that capital punishment would never be justifiable in a democracy. This seems to be a point on which contractualists cannot rightly disagree. Rousseau's assertion that the murderer consents to his own execution may appear to be a counterexample, but Brettschneider maintains that Rousseau's conclusion rests on a bad argument in which Rousseau conflates the distinction he himself had drawn between citizens and persons.[80] Contractualists who reason correctly from contractualist premises therefore will, in Brettschneider's view, necessarily conclude that the death penalty is unjustifiable.

The answer to the first question thus seems to be that contractualism promises as much guidance as should be demanded of a theory of punishment. It will not settle all questions about the proper forms and extent of punishment in detail, but it will settle some of those questions and point the way toward a reasonable range of answers to others. But what of the second question? Must one reason as a contractualist in order to reach Brettschneider's conclusions?

To answer this question, one need only think back to Finkelstein's contractarian approach to punishment. There are differences between her approach and Brettschneider's, to be sure, and I have already indicated that Brettschneider's would not concede as much to criminals as Finkelstein's consent-plus-benefit theory does. Nevertheless, Finkelstein would apparently reach the same conclusion as Brettschneider with regard to wanton violence and limitations on prisoners' rights to free speech and the franchise. She certainly agrees with him in the case of capital punishment. That being so, it seems doubtful that contractualism has a unique contribution to make to the theory of punishment.

To reinforce this point, we need only consider what advocates of other theories of punishment—that is, theories that do not rely on either contractarian or contractualist reasoning—would say about Brettschneider's four claims.

---

[78] *Id.* at 189–90.
[79] *Id.* at 190.
[80] *Id.* at 192.

Utilitarians will certainly agree that it is wrong to employ wanton violence or to inflict pain deliberately "without regard for both the interests of the criminal and the interests of society."[81] Advocates of moral education or restorative justice may well support punitive sentences that allow some significant degree of free speech to prisoners and the vote to those who have completed their sentences.[82] There are even retributivists who will agree that capital punishment ought to be abolished, whether for reasons of the state's fallibility or some conception of human rights or dignity.[83] Where, then, is the value in contractualism?

Brettschneider admits that other theories may be compatible at various points with contractualism.[84] I take it, though, that he would contend that the superiority of contractualism lies not in any particular claim or conclusion it justifies but in the contractualist approach as a whole. For one thing, contractualism rests on a commitment to a reasonable pluralism of views and thus avoids controversial commitments to a comprehensive conception of the good. For another, contractualism brings together a range of judgments that may be compatible with the conclusions that emerge from one or another theory of punishment, but it gives system and coherence to these judgments within the framework of "a distinctly political theory of punishment."[85] By shifting the focus from moral desert to political legitimacy, moreover, contractualism offers the prospect of a properly political response to the two questions with which Morris began his essay: "By what authority do we punish? What permits us to deprive people of their liberty or possessions for some wrong that they have committed?"[86]

These are, in my view, considerable virtues in any theory of punishment. The question is whether they are virtues that only contractualism possesses or even whether contractualism possesses them on its own merits. In pursuing these questions, I shall set aside the point pertaining to contractualism as a theory free from controversial commitments to a comprehensive conception of the good. This is itself a controversial point about which much has been written, including much that maintains that the contractual doctrine of political liberalism is not as neutral among competing conceptions of the good as its adherents claim it to be.[87] But I shall try to show that a theory of punishment grounded in the principle of fair play displays the same virtues as those Brettschneider attributes to contractualism. Beyond that, I shall try to show that whatever virtues contractualism has with

---

[81] *Id.* at 188. For the classical statement of the utilitarian position on punishment, see JEREMY BENTHAM, AN INTRODUCTION TO THE PRINCIPLES OF MORALS AND LEGISLATION 178–88 (1789).

[82] *See, e.g.*, Tim Newell, *Prisoners, Voting, and Active Citizenship*, EKKLESIA (Nov. 18, 2010), http://www.ekklesia.co.uk/node/13601.

[83] *See, e.g.*, Daniel McDermott, *A Retributivist Argument Against Capital Punishment*, 32 J. SOC. PHIL. 317 (2001).

[84] Brettschneider, *supra* note 4, at 182 (referring to "natural law or metaphysical dignity-based understandings of human rights").

[85] *Id.* at 194.

[86] Morris, *supra* note 3, at 53.

[87] *See, e.g.*, MICHAEL J. WHITE, PARTISAN OR NEUTRAL? THE FUTILITY OF PUBLIC POLITICAL THEORY (1997).

regard to punishment are virtues that it owes to its own implicit reliance on considerations of fair play.

V. PUNISHMENT AS FAIR PLAY

The fair-play account is an appropriate standard for assessing contract-based theories of punishment, contractarian as well as contractualist, because it shares an important affinity with them. This affinity is most evident in the way that all three rely on a conception of society as a cooperative venture or enterprise. What Finkelstein says in her essay—"Let us begin with the assumption that society is itself, to use Rawls' phrase, 'a cooperative venture for mutual advantage'"[88]—could also be said by Morris, Brettschneider, and the advocates of the fair-play account. Indeed, Rawls himself provides one of the most influential statements of the fair-play theory of political obligation in an essay that preceded his *A Theory of Justice*:

> Suppose there is a mutually beneficial and just scheme of social cooperation, and that the advantages it yields can only be obtained if everyone, or nearly everyone, cooperates. Suppose further that cooperation requires a certain sacrifice from each person, or at least involves a certain restriction of his liberty. Suppose finally that the benefits produced by cooperation are, up to a certain point, free: that is, the scheme of cooperation is unstable in the sense that if any one person knows that all (or nearly all) of the others will continue to do their part, he will still be able to share a gain from the scheme even if he does not do his part. Under these conditions a person who has accepted the benefits of the scheme is bound by a duty of fair play to do his part and not to take advantage of the free benefit by not cooperating.[89]

Rawls's concern in this essay is with questions of legal and political obligation, but others have taken this principle to provide a justification for punishing those who take advantage of the cooperative sacrifices of others. Their argument runs, briefly, as follows. According to the principle of fair play, anyone who takes part in a cooperative practice and accepts the benefits it provides is obligated to bear his or her share of the burdens of the practice. In the case of the legal order, this means that everyone who profits from others' obedience to the law is under an obligation to reciprocate by obeying the law in turn. Many people, perhaps most, will want to cooperate by obeying the rules of the social or legal order, but they will find it unwise to do so when there is widespread disobedience. In some circumstances, where the sense of community is especially strong, the

---

[88] Finkelstein, *supra* note 3, at 214 (quoting RAWLS, *supra* note 5, at 4).

[89] John Rawls, *Legal Obligation and the Duty of Fair Play*, *in* LAW AND PHILOSOPHY: A SYMPOSIUM 3, 9–10 (Sidney Hook ed., 1964). See RAWLS, *supra* note 5, at 113–14, for Rawls's subsequent doubts about fair play as a grounding for political obligation.

threat of coercion may not be necessary to ensure cooperation. But these circumstances are not likely to prevail in the legal systems of modern states. With the aid of the institution of punishment, however, we can provide a guarantee that "those who would voluntarily obey shall not be sacrificed to those who would not."[90] The practice of punishment is justified, then, because it is necessary to the maintenance of the legal and social order. As long as that order itself is just, or reasonably so, and as long as we cannot trust everyone to obey its rules, we may act to secure its survival by subjecting lawbreakers, who take unfair advantage of the cooperation of others, to punishment.

This sketch of the fair-play account of punishment may prompt questions about its adequacy. One may want to know, for example, what is involved in *taking part in a cooperative practice*, or what it means to *accept the benefits* of such a practice. These and other questions are part of an on-going debate as to the merits of fair play as a justification for punishment, and it is only fair to say that many political and legal philosophers remain to be convinced.[91] The sketch should give some indication of the intuitive power of the fair-play approach, however, and that is enough for present purposes. Those purposes require only enough of an account of the fair-play approach to punishment to make it possible to assess the contractarian and contractualist theories we have been considering in the light of that account.

To begin once again with the contractarians, Morris's argument that people make themselves liable to punishment by forfeiting some or all of their moral standing bears a clear resemblance to the fair-play account. As we have seen, Morris traces the loss of moral standing to the failure to act justly, with justice understood "as a mutually beneficial convention which constrains rational agents in situations where individually rational action leads to disadvantageous outcomes. Compliance with the requirements of justice is the condition for protection by those requirements."[92] But that is to say that the unjust person is one who tries to enjoy the benefits of "a mutually beneficial convention" without bearing its burdens—someone, that is, who tries to take unfair advantage of others. Morris may appeal to rational choice within a hypothetical contract here, but the underlying sentiment is an implicit appeal to the principle of fair play. Furthermore, Morris must trade on considerations of fair play in this regard, for we have already noted that the hypothetical contract will not do the necessary work. He may begin with rational, amoral agents and expect that the contract will somehow transform them into moral actors, but neither he nor others have shown

---

[90] H.L.A. HART, THE CONCEPT OF LAW 193 (1961).

[91] See Richard Dagger, *Playing Fair with Punishment*, 103 ETHICS 473 (1993) [hereinafter Dagger, *Playing Fair*], for indications of the extent and nature of this debate. *See also* Richard Dagger, *Punishment as Fair Play*, 14 RES PUBLICA 259 (2008) [hereinafter Dagger, *Fair Play*]; Davis, *supra* note 2.

[92] Morris, *supra* note 3, at 66.

how this transformation takes place.[93] Fair play slips in between the lines, as it were, to justify the claim that those who act unjustly forfeit some or all of their moral standing.

Much the same is true of Finkelstein's consent-plus-benefit version of contractarianism. She and the fair-play theorists share the view, with Rawls, that society is a collective venture for mutual advantage. They will also agree that those persons who defect from the cooperative venture by breaking its rules will open themselves to punishment, but with the qualification that "defections can be large or small, and it may be that it is still advantageous to cooperate with those responsible for small defections."[94] On these points, though, Finkelstein is simply taking what is fundamentally the fair-play position. She parts company with that position, however, when she argues in terms of hypothetical consent for a system of punishment that provides a net benefit even to the person being punished.[95] There are complications here, but the fair-play position holds that consent—whether express, tacit, or hypothetical—is not a necessary condition for participation in a cooperative practice.[96] There are limits to what can be taken to be a cooperative venture for mutual advantage, certainly, and those unfortunate persons who live under oppressive, exploitative regimes may rightly conclude that they have no obligation grounded in fair play to obey the edicts of their rulers. Consent is not necessary, however, nor is net benefit to everyone in the society, including those who are undergoing punishment. What the fair-play theory will require is that every member of the legal system be treated fairly, as a full member of the cooperative practice who is entitled to an equal consideration of his or her interests. What this amounts to, though, is not a net benefit to everyone in the practice but *a fair chance to enjoy a net benefit* from the practice. Considered simply as a member of the practice, everyone will benefit equally from the existence of the institution of punishment; considered as actual persons, however, all that is necessary is that everyone receive the benefit of law secured by punishment—even if that does not result in a net benefit to him or her.

In this respect, the fair-play account has more in common with contractualism than with contractarianism, but that is because contractualism relies even more on an implicit appeal to considerations of fair play than contractarian theories do. As we have seen, contractualists begin with reasonable persons who do not need a hypothetical contract or rational choice situation to prove to them that it is wrong to enjoy the benefits of a cooperative scheme while shirking its burdens. Rational

---

[93] *See* Lovett, *supra* note 19, at 738 ("Suppose that a group of self-interested and prudential bargainers agree on terms of social justice that would be mutually beneficial if compliance was common. The question remains: What reason does any given member of that group have to comply with those terms on the occasions when unilateral defection would happen to promote self-interest? Within the mutual advantage framework, no workable answer to this last question has yet been supplied.").

[94] Finkelstein, *supra* note 3, at 218.

[95] *Id.* at 215–16.

[96] *See* RICHARD DAGGER, CIVIC VIRTUES: RIGHTS, CITIZENSHIP, AND REPUBLICAN LIBERALISM 72–78 (1997).

agents aim at maximizing their interests, but reasonable persons, as Rawls says, "desire for its own sake a social world in which they, as free and equal, can cooperate with others on terms all can accept."[97] One might also say that reasonable persons desire for its own sake a social world in which everyone plays fair, but failing that perhaps utopian outcome, they will aim to ensure what Rawls calls a fair system of social cooperation over time.

That desire is also reflected in Brettschneider's contractualist theory of punishment. In its general aspect, this requires a punishment scheme that reasonable citizens can justify even "to those who are guilty of crimes."[98] This justification will not aim to show that the convicted criminal will receive a net benefit from the existence of punishment, even after the pain of his or her punishment is taken into account, but it will aim to show that the punitive system is one that he or she can accept as a reasonable citizen. But this is to say that the criminal, *qua* citizen, has no complaint as long as he or she is treated fairly under the rules of the system. Reasonable persons aim to establish a fair system of social cooperation, and reasonable citizens will have no complaint, even when subjected to punishment, if they are treated fairly within such a system. It seems, in short, that it is fairness rather than contractualism that is doing the justificatory work here.

We should turn from the general to the specific, however, to see whether Brettschneider's specific claims are independent of appeals to fairness. If they are, then there will be reason to believe that contractualism is less reliant on considerations of fairness than I have suggested. But if they are not, the case for fair play as the underlying foundation for contractualist punishment will be even stronger.

Brettschneider's first specific claim, again, is that punishments "that deliberately inflict pain without regard for both the interests of the criminal and the interests of society are reasonably rejected."[99] He associates such unjustified punishment with "wanton violence," and particularly with prison guards' "wanton infliction of pain on prisoners."[100] So stated, it is easy to see that reasonable citizens would reject policies that condoned such actions. But what theory of punishment would not? I have already indicated that utilitarians will readily condemn such policies, but that can hardly satisfy Brettschneider, who subscribes to a general moral and political theory, contractualism, that is at odds with utilitarianism. Utilitarians would be joined by advocates of every other theory of punishment in rejecting such policies, however, including those who advocate the fair-play approach. This approach rests on the premise that punishment is necessary to make social cooperation under law possible, and any policy that refused to take account of the interest of the criminal and of society—let alone *wanton* violence—would run counter to that premise.

---

[97] RAWLS, *supra* note 10, at 50.

[98] Brettschneider, *supra* note 4, at 176.

[99] *Id.* at 188.

[100] *Id.*

Fair play is also consistent with Brettschneider's second and third claims, which concern the "democratic rights" of free speech and the franchise.[101] Fair-play theory aims at securing cooperation in the legal order by punishing those who take unfair advantage of law-abiding members of the polity, but it also embraces the aspiration to restore criminals, whenever possible, to the ranks of the law abiding. If an offender has paid his or her debt to society, then fairness requires that society give the ex-offender a chance to resume his or her place among those who participate in the cooperative practice. Whether that aspiration can be achieved by extending rights of free speech and the franchise to criminals, however, does not admit of a straightforward answer. The fair-play theorist will have to balance the interests of social order with those of the wrongdoer, but that is a balancing act, as we have seen, that Brettschneider's approach also must undertake. In both cases the question seems to be, what does fairness require here? Surely criminals retain some of their rights to free speech, but how far should these rights extend while they are imprisoned? Surely fair play requires us to restore voting rights to felons who have served their time—and here fair-play theorists will join Brettschneider in rejecting laws to the contrary—but does that extend to voting rights at all, or of a limited nature, while they are imprisoned? What Brettschneider's reasonable citizens will agree to in these cases seems once again to be what fairness requires.

Brettschneider's fourth specific claim is that reasonable citizens must reject capital punishment, and here there may seem to be a divergence from the fair-play account. Advocates of this account have not been notable for their support of capital punishment, but many people do defend the death penalty on grounds of fairness: These murderers killed, and it is only fair that they be killed, too. But that is to say that fair-play theorists will have to take this sentiment into account, along with other considerations, as they try to decide whether capital punishment is or is not justified, and some of those other considerations will militate against the death penalty. In particular, they will have to give as much weight to the fallibility of the legal system as Brettschneider does, and doing so will raise the question of whether we can ever have assurance that capital punishment will be fairly administered.

Whether Brettschneider's second contractualist reason for opposing capital punishment is equally important for the fair-play theorist is less clear. In this case Brettschneider argues for "the value of never terminating the relationship between citizens and the legitimate state, a relationship that itself is the foundation for the state's authority."[102] I am not persuaded, however, that this is sound reasoning on either contractualist or fair-play grounds. The relationship at the foundation of the state's authority is between *citizens* and the legitimate state, and even if we accept Brettschneider's assertion that citizenship "presupposes the existence of a flesh-

[101] *Id.* at 189–90.
[102] *Id.* at 193.

and-blood person,"[103] I do not see how this leads to the conclusion that no person—no contract killer or tyrant, to take two examples from Morris—can possibly act so *unreasonably* as to forfeit his or her claim to the status of citizen.

But that is to say, once again, that the contractualist and the fair-play theorist seem to be in the same place in this regard. My own view is that the fallibility concern carries the day, so that considerations of fairness render capital punishment unjustifiable. I do not think this reasoning is altogether conclusive, however, even when clothed in contractualist garb. In this case as in the other three specific claims Brettschneider puts forward, the contractualist argument seems to derive whatever plausibility it has from considerations of fair play.

VI. CONCLUSION

In this paper I have examined in some detail three essays that purport to ground the practice of criminal punishment in reasoning that relies in some way on the idea of a social contract. I have done so for both critical and constructive reasons. On the one hand, each of these three essays is worthy of close and critical analysis. Morris, Finkelstein, and Brettschneider make plausible cases for their contractarian or contractualist approaches to punishment. Yet none of them succeeds—or so I have argued—and their lack of success should lead us to look elsewhere for the proper justification of punishment. On the other hand, the fact that all three essays owe much of their plausibility to their implicit reliance on considerations of fair play should lead us to look to the principle of fair play for that justification. There are two further points to note by way of conclusion, however, one in connection with the critical and one with regard to the constructive aspects of this paper.

Concerning criticism, my examination of the essays by Morris, Finkelstein, and Brettschneider should be taken as a complement to a more sweeping criticism of contract-based arguments for social justice that Frank Lovett has recently advanced. According to Lovett, these arguments inevitably encounter

> what might be called the Euthyphro problem—namely, the ambiguity as to whether the claim is supposed to be that the principles selected express the right account of social justice because they (and not others) would be chosen by reasonable people under suitable conditions; or whether the argument is that reasonable people under suitable conditions would choose those principles (and not others) because they express the right account of social justice.[104]

---

[103] *Id.*

[104] Lovett, *supra* note 19, at 738; *see also* Lovett, *supra* note 20, at 84–85.

Lovett extends this criticism of contract-based theories of social justice to similar theories of punishment in a response to Brettschneider's essay.[105] In his interesting reply to Lovett, Brettschneider says that he has "deliberately attempted to sever" the connection between social contract theories and consent in his account of contractualism.[106] He then goes on to say that his view "is based neither in consent nor in a comprehensive conception of the good, but instead in values common to the shared status of all in a democratic polity as free and equal."[107] There is little that is contractual, it seems, in Brettschneider's contractualism. Why not, then, argue directly for a form of punishment that befits the shared status of free and equal citizens in a democratic polity—a form that treats everyone fairly as a cooperating member, or in some cases as non-cooperating members, of the polity? If neither the contract nor consent is doing any real work in the attempt to justify punishment, then it ought to be abandoned in favor of something that will.

Concerning the constructive aspect of my paper, finally, I freely confess that I have not mounted here the kind of defense the fair-play account needs if it is to withstand the objections of its astute critics. I have made some attempts in this direction in other papers, however, as have others who believe fair play offers the best hope for a satisfactory justification of criminal punishment.[108] For now I shall be content if this essay provokes its readers—especially those who have been entranced by contractarian or contractualist approaches—to consider or reconsider the merits of a theory founded on the principle of fair play.

---

[105] Frank Lovett, *Consent and the Legitimacy of Punishment: Response to Brettschneider*, 35 POL. THEORY 806 (2007).

[106] Corey Brettschneider, *Unreasonable Disagreement: Reply to Lovett*, 35 POL. THEORY 811, 811 (2007).

[107] *Id.* at 812.

[108] Dagger, *Fair Play*, *supra* note 91; Dagger, *Playing Fair*, *supra* note 91; Davis, *supra* note 2, at 92–96; GEORGE SHER, *Deserved Punishment Revisited*, *in* APPROXIMATE JUSTICE: STUDIES IN NON-IDEAL THEORY 165 (1997).